# Initialised Eigenlip Estimator for Fast Lip Tracking Using Linear Regression

Simon Lucey, Sridha Sridharan and Vinod Chandran
Speech Research Laboratory, RCSAVT
School of Electrical and Electronic Systems Engineering
Queensland University of Technology
GPO Box 2434, Brisbane QLD 4001, Australia
s.lucey@qut.edu.au, s.sridharan@qut.edu.au and v.chandran@qut.edu.au

## Abstract

*Multimodal speech processing in which visual facial features are jointly processed with audio features is a rapidly advancing field. Lip movements and configurations provide useful information to improve speech and speaker recognition. However the use of this visual information requires accurate and fast lip tracking algorithms. A new technique is outlined that is able to estimate the outer lip contour directly from a given lip intensity image via linear regression. This estimate can be improved by a active shape model that is able to track a speakers lips without requiring time consuming iterative energy minimization techniques. Results of performance are presented against known tracking algorithms using the M2VTS database.*

## 1. Introduction

The tracking of lips in image sequences is a challenging problem in object recognition due to the lack of dominant image features defining the lip contours. Snakes [8], chromatic thresholding [11] and chromatic texture modelling [4] have been used to some success for lip tracking.

Snakes [8] are effective in the lip tracking process but are quite computationally expensive as they require many iterations to fit the lip contour properly. Chromatic features can also be used to segment a speakers lips from the primarily skin background. These techniques have the advantage of being fast as they are simply pixel based and don't require any syntactic information to restrict the lip shape. However, using colour as a feature has several problems. Firstly, the colour representation of a person obtained by a camera is influenced by ambient light and background. Secondly, different cameras produce significantly different colour values, even for the same person under the same lighting condition [12]. These variabilities make lip tracking through chromatic features problematic.

Active shape models (ASM) have proved to be traditionally very good at providing a model for the deformation of lip contours and in turn provide an accurate way of tracking a speakers lips [5]. ASMs have the advantage of providing a priori knowledge about typical deformation of lips from a training set of labeled lips and are able to be used effectively with gray scale images negating any problems with colour constancy. They are also able to model the high order statistics of lip shape and intensity through an effective eigen decomposition based on principal component analysis (PCA). Previous implementations of ASMs have required computationally expensive algorithms to minimise a cost function due to problems with inaccurate initial lip shape estimates. Our work has refined this technique to first gain an accurate initial estimate of the outer lip contour directly from a given lip intensity image via linear regression. Using this estimate we demonstrate it is possible to gain an accurate lip contour expediently by taking profiles around the estimated lip contour updating the ASM only once.

## 2. Principal Component Analysis

The techniques outlined in this paper rely heavily on PCA to extract features from different feature spaces as well as providing a method for mapping from one feature space to another. So not to confuse between the different feature spaces being used a generalised nomenclature for PCA shall be defined. Generally a feature space (eg. lip intensity image features, lip contour features and lip intensity profiles normal to the contour) can be approximated by,

$$\mathbf{y} \approx \overline{\mathbf{y}} + \mathbf{P}\mathbf{b} \tag{1}$$

where $\overline{\mathbf{y}}$ is the mean of the training feature vectors, $\mathbf{P}$ the matrix of the first few column eigenvectors of the covariance matrix which correspond to the largest eigenvalues and

0-7695-0750-6/00 $10.00 © 2000 IEEE

$\mathbf{b}$ a vector containing the weights for each eigenvector. The vector $\mathbf{b}$ can be used as a compact and decorrelated approximate representation of the original vector $\mathbf{y}$ in which the main modes of variation have been preserved.

## 3. Eigenlips

Eigenlips were first presented by Bregler and Konig [1] in which cropped lip images were decomposed using PCA into lower dimensional subspace for the purposes of speech recognition. The term *Eigenlips* refers to the first $n$ principal components of a gray-level matrix centered and scaled around the lips and is an extension of the *Eigenfaces* work first developed in [10] which dealt with cropped intensity images of the face.

Bregler and Konig [1] demonstrated that the lip ROI intensity image has important lip information contained in it. They showed that since the window around the lip ROI doesn't deform with the lips, the principle modes of variation are mainly attributed to lighting, skin shade and shape variations. In their experiments they demonstrated that the Eigenlip features used directly in lip reading applications performed well but were highly affected by lighting differences and never quite out performed a contour model of the lips.

Instead of using the lip intensity image directly our work has concentrated on a variant by trying to use the lip intensity image to directly estimate the outer lip contour. In our technique the scale of the face and thus to a certain extent the lips is known a priori due to the speakers eyes being tracked a priori. For a input image at any given scale the face and eye locations for a speaker are found using a standardised tracking algorithm employing a probabilistic eigen-template matching technique as described in [6]. The technique tracks objects based on a number of two-dimensional convolution operations which can be computed efficiently using a radix-2 based FFT thus lending it well for real time application. Using the distance between the eyes as a reference of scale the face is cropped and re-scaled to a specific size. Using the same technique employed for the face and eye tracking the lip region of interest (ROI) is found by searching the lower half of the face ROI using Eigenlip template matching.

A random set of images taken from the M2VTS database [7] using 105 images taken across 35 speakers with the 3 shots of each individual speaker was used as our training set. For each of the 105 test images the re-scaled lip ROI's were found for training. PCA was performed on the set of lip training images with each image being statistically normalised before hand as recommended in [6, 9].

The localised $M$ by $M$ lip (ROI) can be expressed as a $M^2$ vector $\mathbf{i}$. Any lip intensity image $\mathbf{i}$, where $\mathbf{i}$ contains the intensity information of the lips, can then be decomposed using PCA so as to extract the principal modes of intensity variation resulting in a feature vector of weights $\mathbf{b_i}$ as modeled in Equation 1. These modes of variation can be attributed to a number of characteristics such as lighting and more importantly lip shape.

### 3.1. Eigencontour

Each image in the training set of 105 cropped lip images had the outer labial lip contour labeled. PCA was performed so as to extract the principal modes of contour shape variation resulting in the feature vector of weights $\mathbf{b_s}$ as modeled in Equation 1.

### 3.2. Using linear regression to estimate Eigencontour

We use linear regression method because of its simplicity for the problem of lip shape estimation. Consider the feature vectors taken from the Eigenlip decomposition $\mathbf{b_i}$ of the lip ROI from a set of test images as the concatenated matrix $\mathbf{V_I}$. Next consider the feature vectors created from PCA analysis on the manually fitted lip contours $\mathbf{b_s}$ of the same set of test images as the concatenated matrix $\mathbf{V_S}$. To get an estimate of lip shape from the given lip intensity image we can calculate the matrix $\mathbf{M}$ that realizes the linear regression between $\mathbf{V_I}$ and $\mathbf{V_S}$ [3]. Finally for every new lip ROI the Eigenlip decomposition feature vector $\mathbf{b_i}$ can be multiplied by the regression matrix $\mathbf{M}$ to obtain an estimate of the lip shape $\mathbf{b_s}$ for that speaker.

This technique works well with the linear regression matrix able to estimate the shape and position of a speakers lips to within a couple of pixels for the majority of lip shapes. Some results can be seen in Figure 1.

These images were part of a testing database separate to the training database used to create the linear regression matrix $\mathbf{M}$. In Figure 1 (a)-(d) the contours track the lip particularly well with the correct placement of a contour around the speakers lips to within a few pixels of the actual lip contour. Images (e)-(i) highlight some problems with the technique as the contours although giving a rough estimate of the lip shape don't fit the actual lip contour properly. Images (h) and (i) were actually taken from our own database to test the technique on different speakers under different lighting conditions not available in the M2VTS database.

### 3.3. Refining the lip estimate

The results of the regression process demonstrated that the technique had some merit but needed refinement to improve the estimated contour. As with the ASMs used in [5] grayscale profiles normal to the lip contour were used to refine the initial lip shape estimate provided by the linear

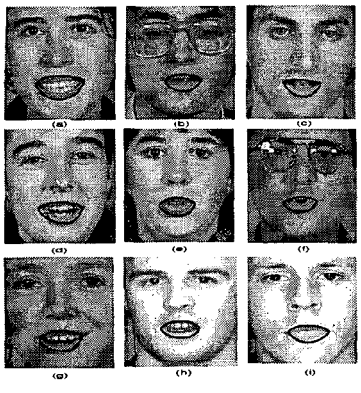

**Figure 1. Example of lip contour estimates found via linear regression (note (h) and (i) were taken under different lighting conditions).**

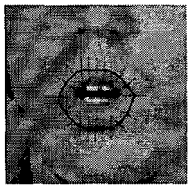

**Figure 2. Intensity profiles taken around contour of lips.**

regression step. This is an essential step as the linear regression estimate needs to be refined so as to be accurately placed and deformed to the speakers lips. We sample one-dimensional intensity profiles perpendicular to the contour at each model point and for each training image as can be demonstrated in Figure 2. The profiles of all model points of a training image are concatenated to form a global profile vector $\mathbf{g}$. The principal modes of intensity variation normal to the contour were then obtained using PCA resulting in the feature vector of weights $\mathbf{b_g}$. This intensity model is used to search profiles taken from the lip estimate to find the best lip contour match along the profile. The best match was found using a probabilistic approach where the distance in feature space and distance from feature space metrics were combined to create an accurate similarity measure [6]. A thorough rundown of this technique can be found in Cootes, Taylor and Haslam [2].

## 4. Comparison of algorithms

An inherent problem with lip tracking is in getting an accurate measure of how well a given algorithm tracks a speakers lips. Luettin [5] has recently proposed that an accurate measure of the quality of a lip tracking algorithm is indicative of how well it performs in a given lip reading application (ie. the recognition of visemes or words). To this end we have constructed a simple isolated word recognition test using just visual lip features so as to gauge the effectiveness of our system.

A hidden markov model (HMM) was trained for ten French digits uttered in the M2VTS database. Six speakers were used in the testing and training phases with each speaker uttering the digits on four separate occasions. Due to the small size of the training/testing set recognition tests were performed using the 'leave-one-out' method ie. Five subjects used for training and one for testing. The whole procedure was repeated six times. The actual word recognition system was made very simple so as to gauge the effectiveness of the lip features and not the recognition system itself. A left to right single mixture 5 state HMM was used for the word models.

Table 1 summarises the results for our digit recognition experiments. The system first accepted speech features (ie. 12 mel-cepstrum coefficients for every 40 ms frame) so as to check effectiveness of the HMM framework in the given digit recognition task and to give a optimal lower benchmark for the results. Visual features were then used to check the effectiveness of our algorithm. Using just the Eigenlip features gave poor accuracy rates. The mapped Eigenlip features into Eigencontour space through linear regression gave a marked improvement in accuracy. Finally the improved Eigencontour using profiles gave a further improvement in recognition. For completeness we compared these results against other tracking algorithms namely the chromatic thresholding technique by Wark and Sridharan [11] and the chromatic B-splines technique by Sanchez, Matas and Kittler [8]. These algorithms were used as they were freely available for comparison on the M2VTS database and provided contour information for the outer labial contour like our algorithm. As can be seen in Table 1 our final Eigencontour out performs both those algorithms by a clear margin.

## 5. Conclusions

The mapping from Eigenlip intensity space to Eigencontour space was able to improve word recognition results over the intensity features through a simple linear transformation. This linear transformation can be viewed as an emphasizing filter that is able to emphasize aspects of the intensity feature space (eg. outer labial contour) that aid

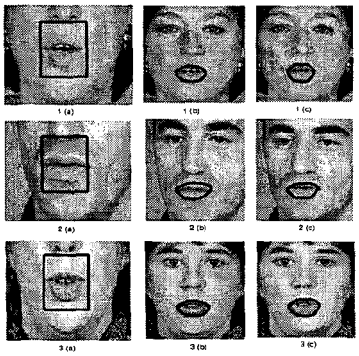

**Figure 3. Demonstration of lip tracking process with the initial tracking (a) then estimate of lip shape through linear regression (b) finally updating estimate with gray scale profiles (c).**

| Features | Recognition (%) |
|---|---|
| mel cepstral* | 0 |
| Eigenlip $\mathbf{b}_i$ | 48.94 |
| estimated Eigencontour via linear regression $\mathbf{b}_s$ | 59.22 |
| improved Eigencontour using profiles $\mathbf{b}_s$ | 65.23 |
| *Wark and Sridharan [12]* R/G Threshholding | 54.34 |
| *Sanchez, Matas and Kittler [9]* B-splines | 48.09 |

**Table 1. Lip reading results where both static and delta features were used in all tests (note * used speech features for testing word recognition engine).**

in the lip reading process. Word recognition results could also be improved by estimating the inner labial contour to increase the visual features being used for the task of lip reading. Updating the estimated profile by taking profiles around the contour improved recognition rates but also increased computation rates when compared to chromatic thresholding techniques [11] and the Eigenlip techniques. Further work could deal with improving the contour estimation through a non-linear mapping procedure that could give accurate recognition results without the added computation associated with deforming the contour via profiles. The results for the word recognition experiments would have most likely been improved if temporal constraints were placed on the contours being tested instead of extracting the contour based purely on a single frame. Results could of been further improved by using multiple mixtures in the HMMs and using data from the inner labial contour as well as the outer labial contour. The estimation through linear regression of the outer labial contour has shown to be affected by changes in lighting conditions. Our future work will deal with making the estimation process more robust for varying lighting conditions while improving the accuracy of the estimated contour.

A new technique has been presented for tracking a speakers lips that is able to extract features useful for the task of lip reading. The technique has also shown to be computationally feasible when compared to the quick chromatic lip trackers. The authors would like to thank the M2VTS project for the use of their database.

## References


[1] C. Bregler and Y. Konig. Eigenlips for robust speech recognition. In *Proc. of IEEE Int. Conf. on Accoustics, Speech and Signal Processing*, Adelaide, Australia, 1994.

[2] T. Cootes, A. Hill, C. Taylor, and J. Haslam. Use of active shape models for locating structures in medical images. *Image and Vision Computing*, 12(6):355–365, August 1994.

[3] E. Kreyszig. *Advanced engineering mathematics*. John Wiley and Sons, Inc., 7th edition, 1993.

[4] M. Lievin and F. Luthon. Unsupervised Lip Segmentation under Natural Conditions. In *ICASSP' 99*, pages 3065–3068, Phoenix, Arizona, March 1999.

[5] J. Luettin, N. Thacker, and S. Beet. Statistical lip modelling for visual speech recognition. In *VIII European Signal Processing Conference*, Trieste, Italy, 1996.

[6] B. Moghaddam and A. Pentland. Probabilistic Visual Learning for Object Representation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7), July 1997.

[7] S. Pigeon. The M2VTS database. Laboratoire de Telecommunications et Teledection, Place du Levant, 2-B-1348 Louvain-La-Neuve, Belgium, 1996.

[8] M. U. Ramos Sanchez, J. Matas, and J. Kittler. Statistical chromaticity models for lip tracking B-splines. In *Int. Conf. Audio and Video based Biometric Person Authentication*, Crans Montana, Switzerland, 1997.

[9] M. Turk and A. Pentland. Eigenfaces for Recognition. *J. Cognitive Neuroscience*, 3(1), 1991.

[10] M. Turk and A. Pentland. Face Recognition Using Eigenfaces. 1991.

[11] T. Wark and S. Sridharan. A syntatic approach to automatic lip feature extraction for speaker identification. In *ICASSP' 98*, pages 3693–3696, May 1998.

[12] J. Yang and A. Waibel. A Real-Time Face Tracker. In *Proceedings of WACV'96*, pages 142–147, Sarasota, Florida, USA, 1996.